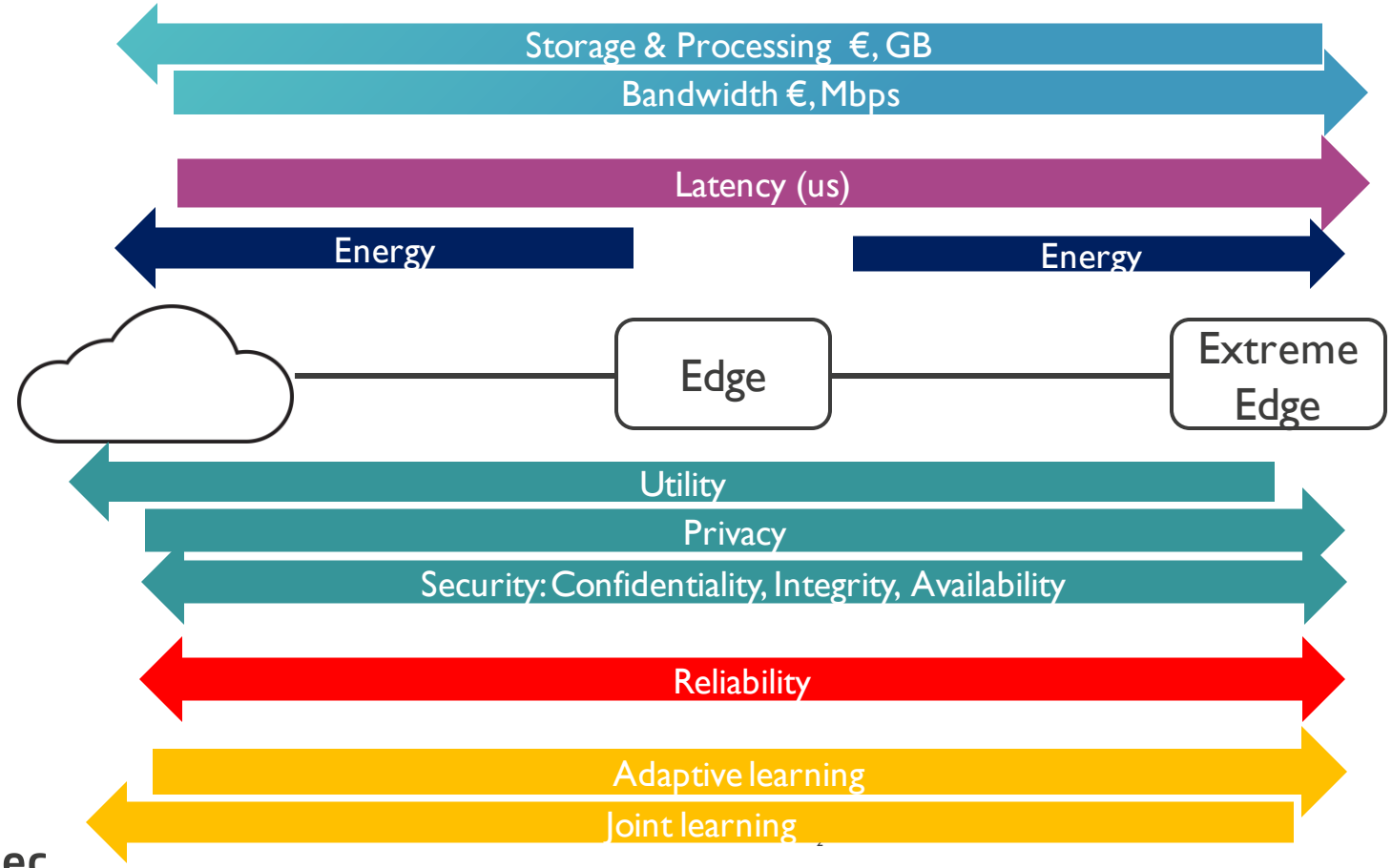
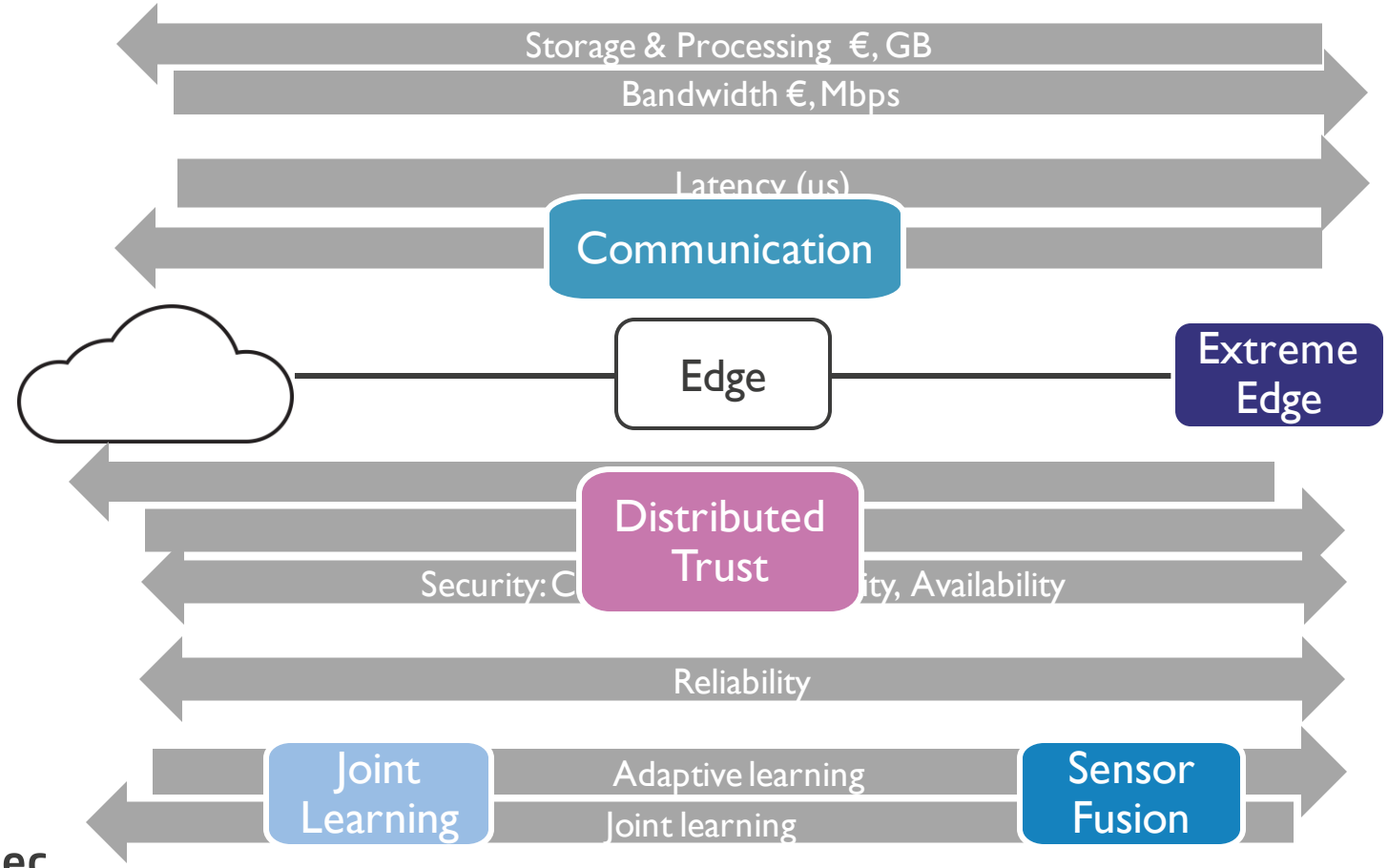
LIVING AT THE EXTREME EDGE: TOWARDS DISTRIBUTED HIGH PERFORMANCE ARTIFICIAL INTELLIGENCE

RUDY LAUWEREINS – VICE PRESIDENT IMEC & PROFESSOR KU LEUVEN

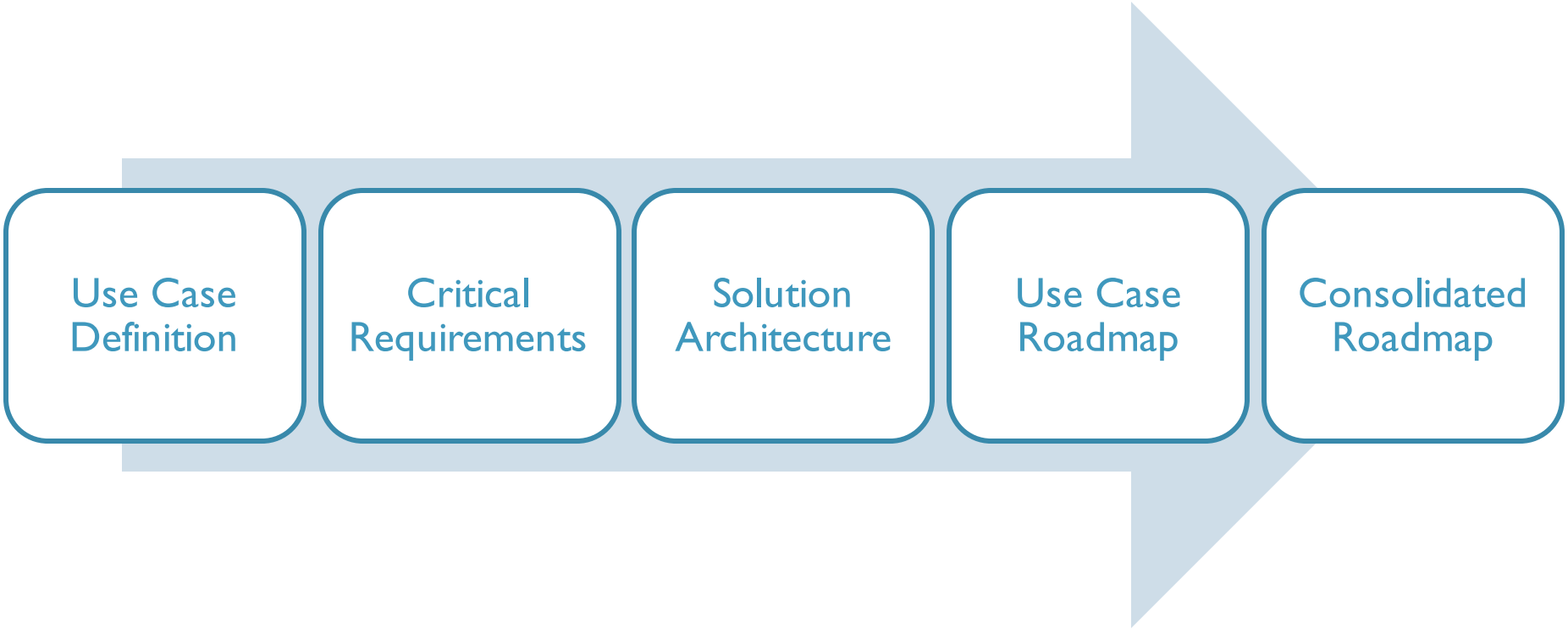
THINK DISTRIBUTED…



5 FOCUS AREAS



A CASE DRIVEN APPROACH



SUMMARY OF DISTRIBUTED AI USE CASES AND REQUIREMENTS

	Security	Mobility	Sequencing	AR	Industrial IoT
DESCRIPTION	Biometric authentication using gait	Supervised intersections	Sequencing of cancer cells in the blood	From single user to real-time, collaborative and perceptive AR	Smart factory quickly reconfiguring (c/r) obot production line
STORAGE/MEMORY			■		
PROCESSING	■			■	
COMMUNICATION		■	■		
LATENCY		■		■	■
POWER CONSUMPTION	■	■	■	■	
PRIVACY	■				
SECURITY	■				■
RELIABILITY & AVAILABILITY		■	■		■
ADAPTIVE LEARNING	■	■			
JOINT LEARNING		■	■		

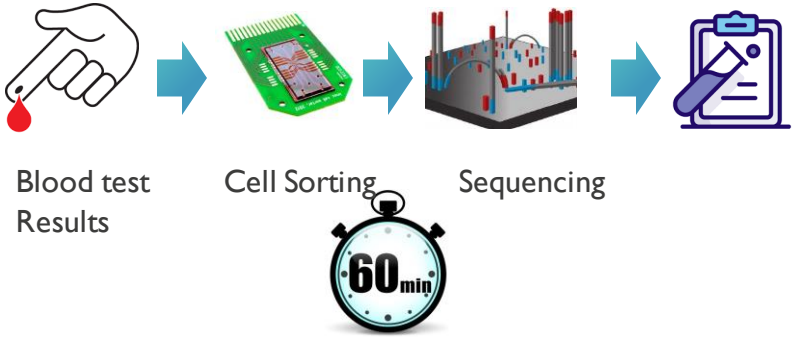
imec

ADDITIONAL USE CASES

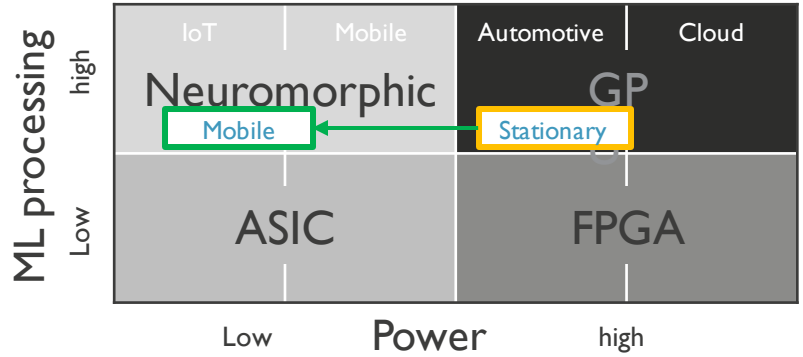
DISTRIBUTED & COLLABORATIVE AI

CASE STUDY: GENERIC SCREENING PLATFORM FOR BLOOD OR WATER

From drop of blood to high-quality analysis within 1 hour



Virus, Cancer, RNA	60 min
10M DNA fragments, 800K fps	2 TB/s
< $300	80 TOP/s

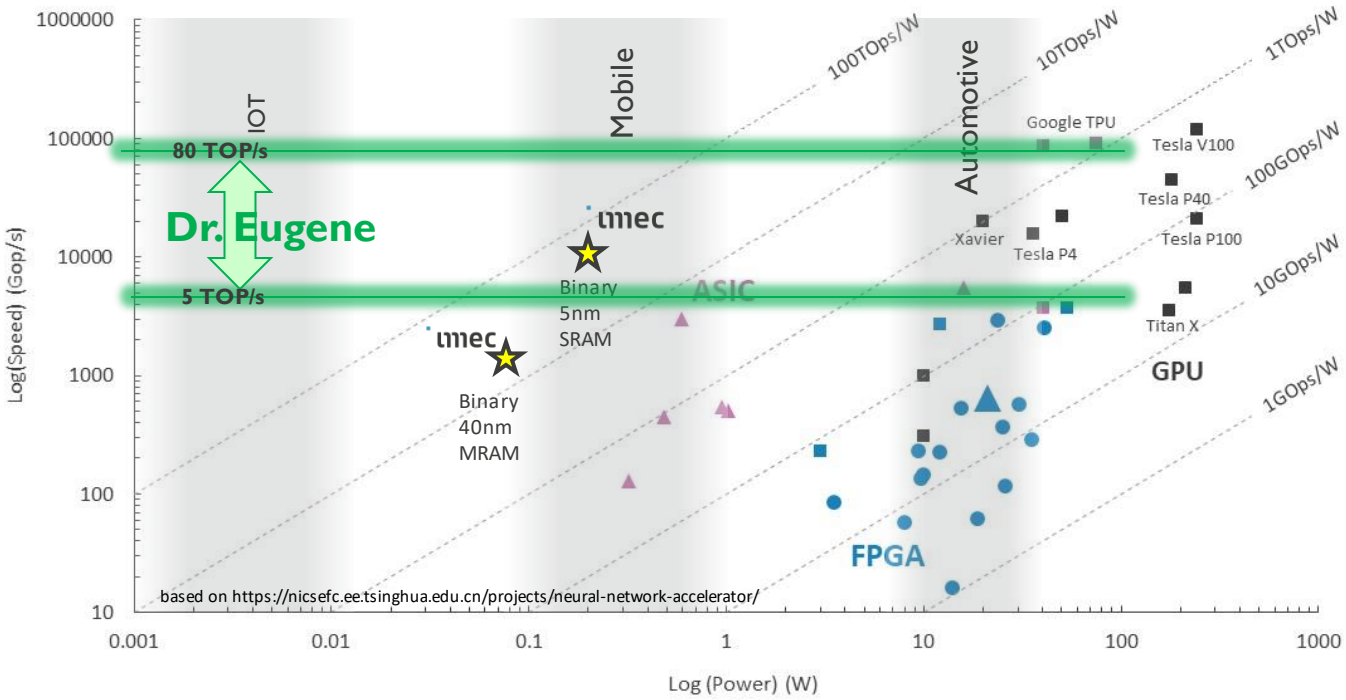


Generic Screening Platform for Blood or Water

- Match against Virus Fingerprints, Cancer Fingerprints, RNA Sequencing, Chronical Oncology Follow-up
- Water: Match against Virus Fingerprints
- Battery powered devices needed for field operations in e.g., emergency care (blood), field operations (blood and water), the latter specifically in developing countries and for military use.

HIGH ENERGY EFFICIENCY NEURAL NETWORK HARDWARE

BENCHMARK – DR. EUGENE TARGETS



DHPAI PROGRAM

HEALTH – DR.EUGENE USE CASES

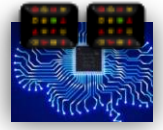
	1-2 years	2-5years	5+ years
Sensor Fusion	**Combining sensors** Training models end-to-end fusing multiple sensor inputs on single device	**Distributed sensors** Sensor information coming from devices distributed in the network **Robust sensor fusion** Robust against sensors added / removed / failing	**High performance A.I.** Optimized A.I. models for optimized hardware **Transfer learning** Transfer learned models to a new instance of the system: - new environment - other sensors on different locations - different agents - …
Joint Learning	**Combining data sources between learning agents** - Learn locally on each device using it's own data source - Improve by sharing information, without sharing the whole local dataset	**Robust joint learning** Robust against agents being added/removed to the system	**World modeling** A.I. agents that understand the world they operate in - process multiple sensors inputs - enable to plan ahead - multi-agent interactions - zero-shot learning through "imagination" - …

Applications

Virus Scanner

Oncology Scanner

Field Updateable Scanners

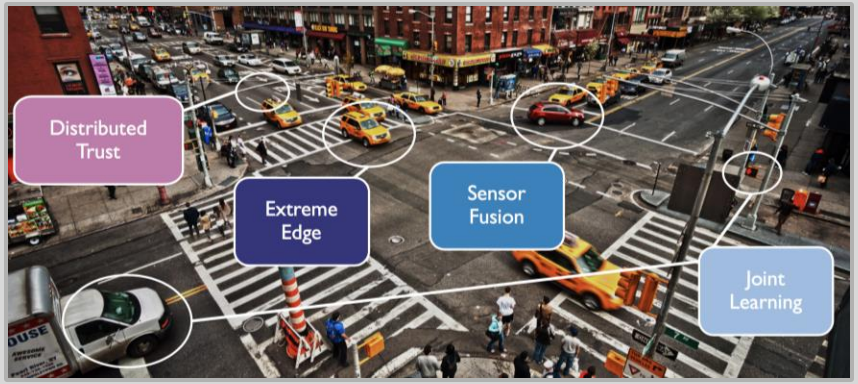


Field Adaptive Scanners

Oncology Treatment Decision Support

MOBILITY: SUPER-VISION

From autonomous traffic to supervised intersections



Up to 100 actors

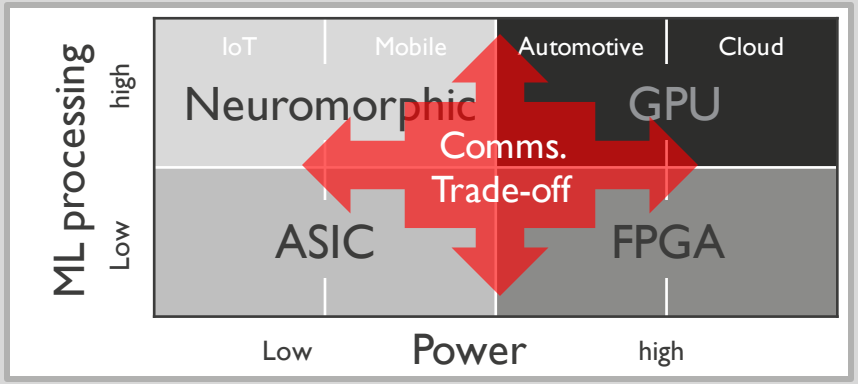
Data streaming easily tops 1Gbps

Realtime awareness requires latencies <20ms

Supervised navigation of the intersection requires latencies <10ms

Significant work needed to define interchange formats

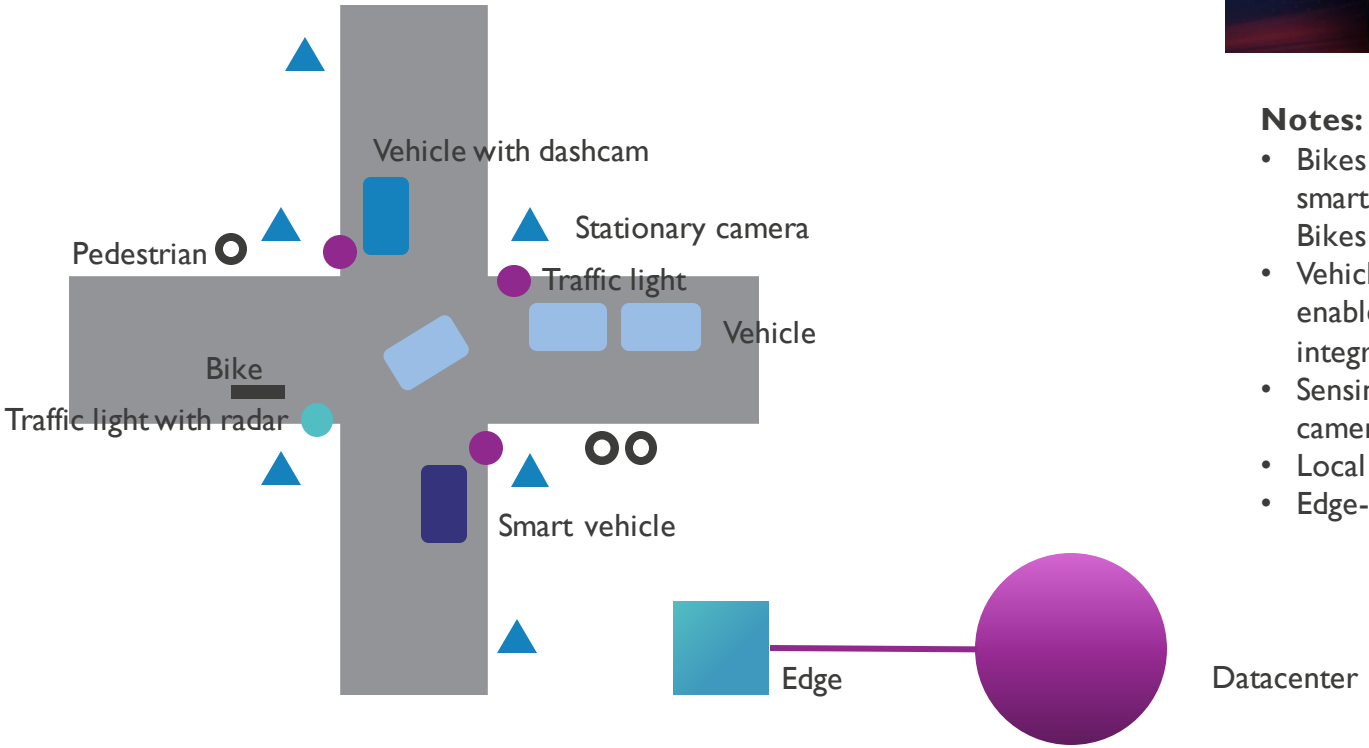
Data streaming can dominate power budget



- Smart intersections that evolve from stupid ones (monitoring) to smart ones (data exchange between actors) to supervised ones (intersection controls the cars when they enter its domain).
- Model scaling implies gradually growing amount of "smartness" i.e. data captured and processed by the traffic.
- Neither on device, nor edge or cloud provides a closed solution in our first-order analysis
- Open questions remain on the UI/UX aspects of interactions with bikes, pedestrians (unless we just move them to dedicated lanes and bridges).

SUPER-VISION
ARCHITECTURE DEFINITION

Canonical intersection

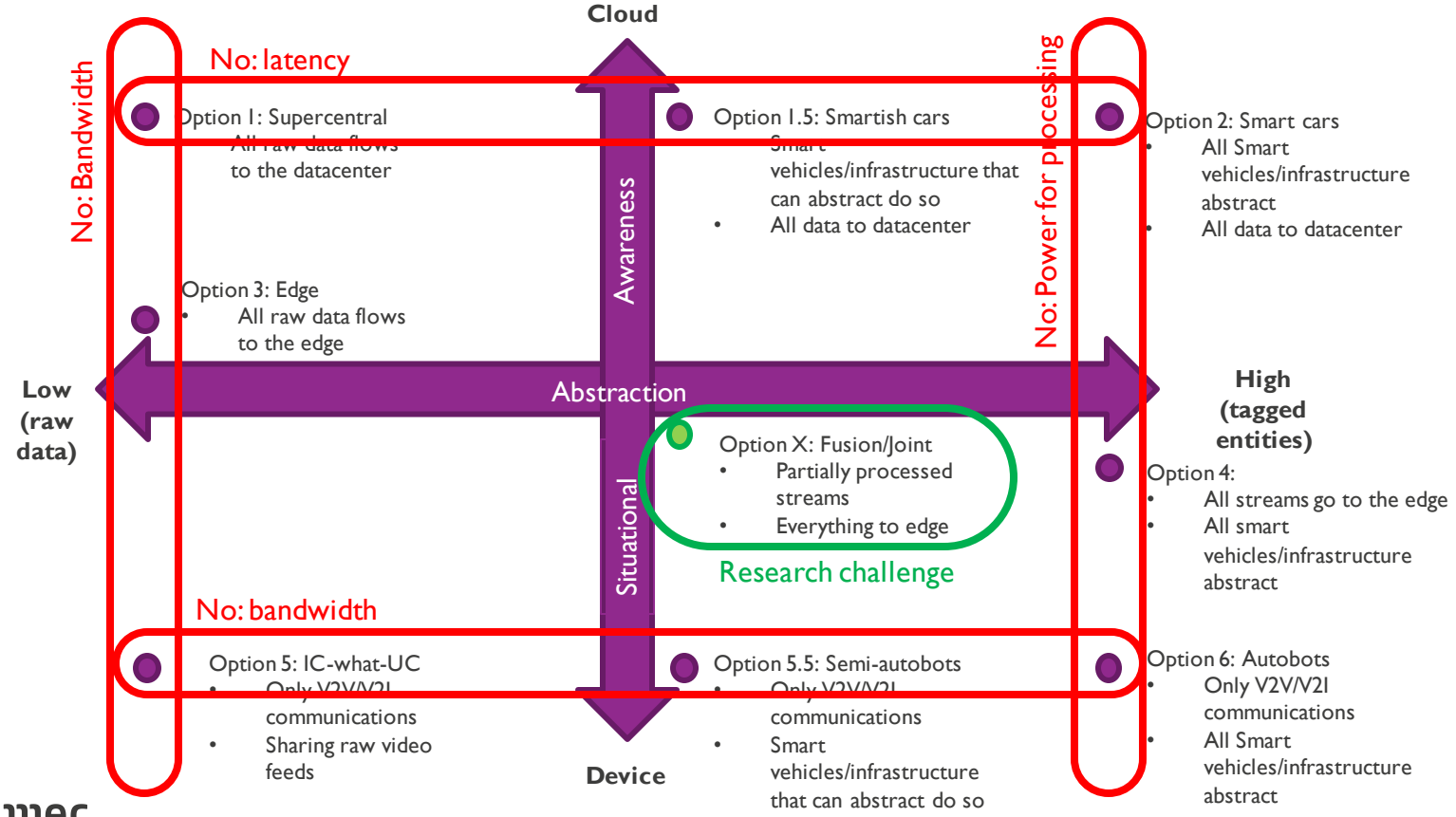


Notes:

- Bikes and pedestrians can be smartphone-enabled.
  Bikes can be e-bikes
- Vehicles that are smartphone-enabled can have several levels of integration.
- Sensing vehicles (L2+) have multiple cameras and radars
- Local edge connects wirelessly
- Edge-datacenter is wired

ARCHITECTURE OPTIONS

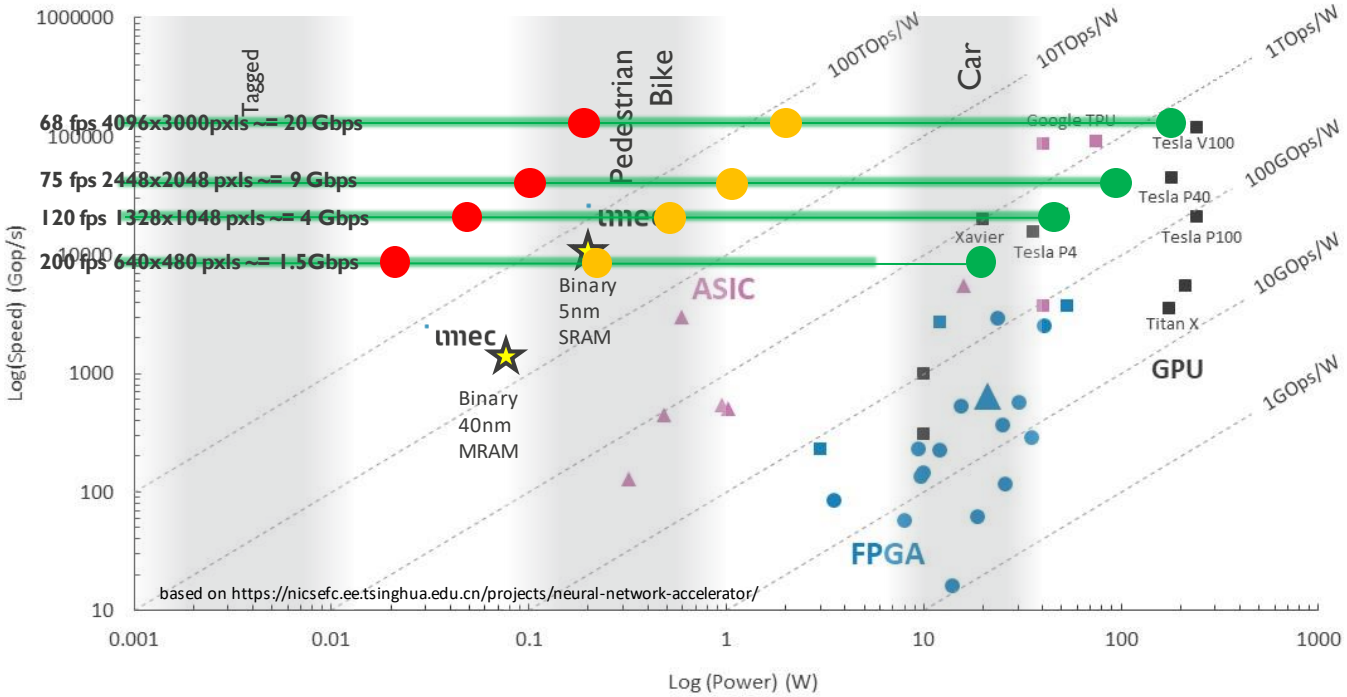
DEPENDING ON LOCAL PROCESSING AND AWARENESS



imec

TRADING OFF PROCESSING WITH TRANSMISSION (QUAD*)

BENCHMARK – TRANSMISSION OF IMAGE DATA VS PROCESSING



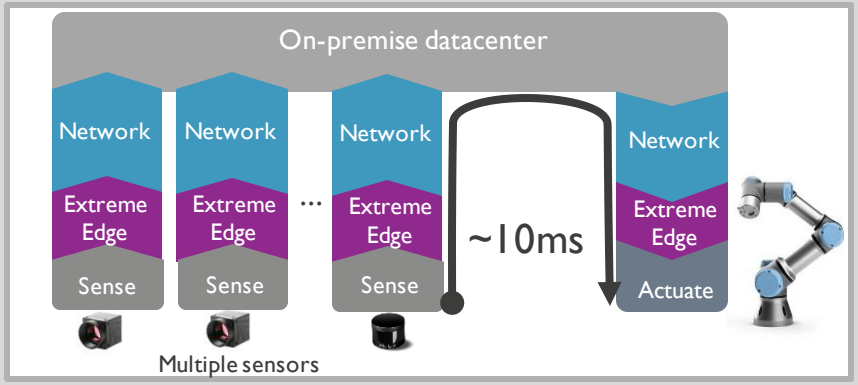
- Power requirement only to transmit the raw data (lowest latency), where we're using an (optimistic) 10^{-8} J/bit average number.
  As can be seen, moving all processing to a central cloud can consume almost as much energy as the processing itself.
  Compression is equivalent to (partial) local processing with one key difference: **latency**. ●, ● are factor 1000 (HEVC target), 100

imec

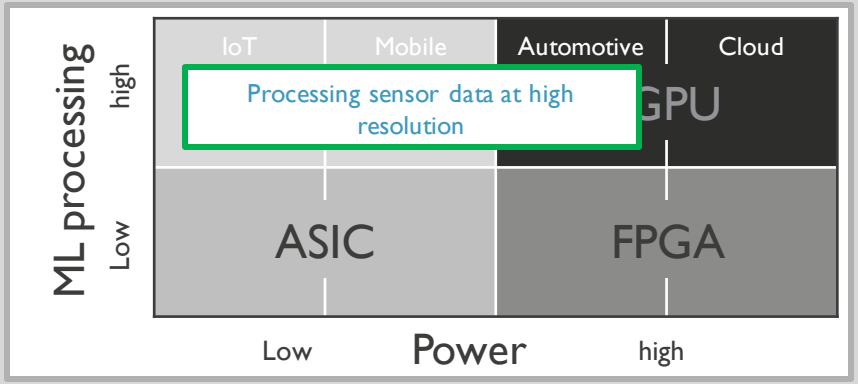
*QUAD = quick-and-dirty

INDUSTRIAL IOT

Closed-loop sensor-based control for robotic manipulators



~10 ms latency	Fuse multiple sensors on-premise
High resolution and framerate	Joint learning from many robots
10-100 TOPs	>99,9999 availability



- Closed-loop robot control requires low latency processing (~10ms) of high resolution sensors (i.e. camera, LIDAR, …).
- Currently, processing one sensor on a high-end GPU is feasible, but multiple sensors create an I/O bottleneck.
- Neuromorphic hardware stacked on the sensing/actuating chip ("extreme edge") could help to:
  - Reduce the latency and bandwidth to the fusing A.I.
  - Scale number of sensors as each sensor has dedicated processing power
  - Increase the number of actuators enabling joint learning

LATENCY REQUIREMENTS FOR INDUSTRY 4.0

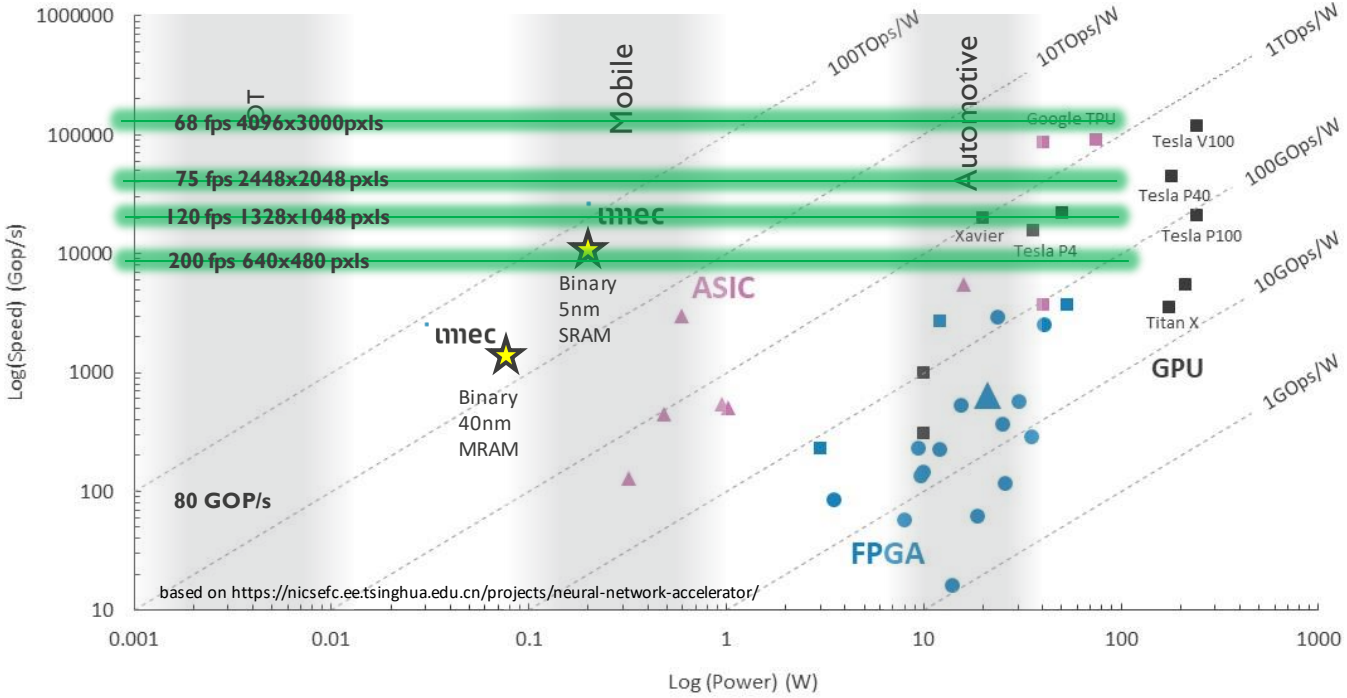
SOURCE: 3GPP 22.804

Use case (high level)		Availability	Cycle time	Typical payload size	# of devices	Typical service area
Motion control	Printing machine	>99.9999%	< 2 ms	20 bytes	>100	100 m x 100 m x 30 m
	Machine tool	>99.9999%	< 0.5 ms	50 bytes	~20	15 m x 15 m x 3 m
	Packaging machine	>99.9999%	< 1 ms	40 bytes	~50	10 m x 5 m x 3 m
Mobile robots	Cooperative motion control	>99.9999%	1 ms	40-250 bytes	100	< 1 km^2
	Video-operated remote control	>99.9999%	10 – 100 ms	15 – 150 kbytes	100	< 1 km^2
Mobile control panels with safety functions	Assembly robots or milling machines	>99.9999%	4-8 ms	40-250 bytes	4	10 m x 10 m
	Mobile cranes	>99.9999%	12 ms	40-250 bytes	2	40 m x 60 m
Process automation (process monitoring)		>99.99%	> 50 ms	Varies	10000 devices per km^2	

imec

HIGH ENERGY EFFICIENCY NEURAL NETWORK HARDWARE

BENCHMARK – PROCESSING IMAGES OF ONE SINGLE CAMERA



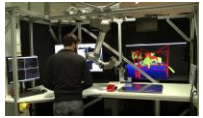
DHPAI PROGRAM

INDUSTRIAL IOT – ROBOTICS USE CASES

	1-2 years	2-5years	5+ years
Sensor Fusion	**Combining sensors** Training models end-to-end fusing multiple sensor inputs on single device	**Distributed sensors** Sensor information coming from devices distributed in the network **Robust sensor fusion** Robust against sensors added / removed / failing	**High performance A.I.** Optimized A.I. models for optimized hardware **Transfer learning** Transfer learned models to a new instance of the system: - new environment - other sensors on different locations - different agents - …
Joint Learning	**Combining data sources between learning agents** - Learn locally on each device using it's own data source - Improve by sharing information, without sharing the whole local dataset	**Robust joint learning** Robust against agents being added/removed to the system	**World modeling** A.I. agents that understand the world they operate in - process multiple sensors inputs - enable to plan ahead - multi-agent interactions - zero-shot learning through "imagination" - …

Applications

Multi-camera quality inspection on assembly line



Multi-sensor workcell tracking

Collaborative drones

Flexible assembly line (re)configuration

Real-time, closed-loop robot control

imec

embracing a better life